Lattice-based lightly-supervised acoustic model training

Lattice-based lightly-supervised acoustic model training

Joachim Fainberg, Ondřej Klejch, Steve Renals, Peter Bell

Centre for Speech Technology Research, University of Edinburgh, United Kingdom

{j.fainberg, o.klejch, s.renals, peter.bell}@ed.ac.uk

Abstract

In the broadcast domain there is an abundance of related text data and partial transcriptions, such as closed captions and subtitles. This text data can be used for lightly supervised training, in which text matching the audio is selected using an existing speech recognition model. Current approaches to light supervision typically filter the data based on matching error rates between the transcriptions and biased decoding hypotheses. In contrast, semi-supervised training does not require matching text data, instead generating a hypothesis using a background language model. State-of-the-art semi-supervised training uses lattice-based supervision with the lattice-free MMI (LF-MMI) objective function. We propose a technique to combine inaccurate transcriptions with the lattices generated for semi-supervised training, thus preserving uncertainty in the lattice where appropriate. We demonstrate that this combined approach reduces the expected error rates over the lattices, and reduces the word error rate (WER) on a broadcast task.

Index Terms: Automatic speech recognition, lightly supervised training, LF-MMI, broadcast media

1. Introduction

Automatic Speech Recognition (ASR) systems are ideally trained on accurate transcripts that match the audio. There is, however, a large amount of available data that has inaccurate and partial transcripts. Examples include traditional broadcast data [1, 2, 3, 4], YouTube data [5], medical data [6], crowd-sourced data [7], and children's data [8].

Obtaining manual transcriptions is expensive. Instead, there is a growing body of work on lightly supervised methods that aim to make use of partial and inaccurate transcriptions for aligning and training acoustic models. Most of these methods make use of a decode of the data made with a language model (LM) biased towards the data itself [9, 10]. They typically proceed by either filtering the data at various levels of granularity [1, 5, 11, 2], or by some combination or error correction algorithm [4, 12, 13, 14, 7, 8, 15, 16]. The benefit of the latter techniques is that they can often maintain more data than through filtering, while filtering is perhaps most appropriate with large amounts of data such that false rejects is a non-issue. New acoustic models can be trained on the filtered or corrected transcriptions, and sometimes the process is repeated, yielding increasingly improved models [17, 8].

In contrast to the lightly supervised approaches, semi-supervised training requires no transcriptions for new data, but generates hypotheses using a large (non-biased) background model. In state-of-the-art approaches using the discriminative LF-MMI objective (see [18]), the decoding lattices are maintained and used as lattice supervision, effectively encoding the uncertainty of the hypotheses by the width of the lattice [19, 20]. As the authors remark, this is beneficial for discriminative training which is sensitive to the accuracy of the supervision [6, 21].

The contribution of this paper is three-fold. First, typical lightly supervised techniques produce single best path transcriptions on which to train. Yet, state-of-the-art semi-supervised, discriminative, training techniques benefit strongly from lattice supervision which encodes the uncertainty of the data [19, 20]. We experiment with lightly-supervised training where the lattice supervision is generated with a biased LM, and demonstrate that this can substantially improve WERs.

Second, Long et al. [4] showed that, instead of filtering, it is possible to combine inaccurate transcripts with a biased decode lattice to create an improved best path transcription on which to train. However, the output is a best path, not a lattice. Manohar et al. [16] demonstrated an algorithm that combines individual transcripts into a confusion network lattice for training, though this approach does not combine with a hypothesis lattice. We propose a new method to combine the transcriptions, and a hypothesis lattice, while, crucially, maintaining a lattice for supervision. This encodes uncertainty where the transcriptions and lattices disagree, whilst maintaining a narrow lattice where they do agree. We show up to 17.5% relative reductions in WER with respect to a semi-supervised baseline.

Finally, in our experiments we observed that the proposed method compensates for a large number of deletions. We propose to reduce deletions by rewarding insertions when *generating the supervision lattices*, which in our experiments reduces WERs up to 13% relative. The combined improvements from the above ideas yield up to 20% relative WER reductions on broadcast data from the Scottish Parliament, adapting from a model trained on BBC news broadcasts.

2. Related work

Lamel et al. [9] introduced filtering based on segment-level matching with biased LMs. This approach was later extended to discriminative training [10]. Typically segments are filtered given a matching error rate threshold at the word (WMER) or phone level (PMER) between the transcriptions and the biased decode [1, 22, 23, 4]. Methods operating at finer levels of granularity often include selecting *islands* of consecutive words with zero string edit distance [24, 5, 2, 6], or to select words based on a set of cascaded classifiers [25]. Some approaches consider the alignment and match of two transducers, one which allows word-skips (*skip-net*), and one which doesn't (*sequence net*) [17, 2], or use a factor transducer [26] to select reliable segments. Most approaches can be iterated, yielding more and better data with increasingly refined models.

Combination approaches, on the other hand, aim to maintain as much data as possible through correcting or combining the hypotheses with the transcriptions. Long et al. [4] proposed a word-level combination scheme that uses ROVER [27] to select a sequence of words from reference transcriptions with the corresponding hypothesis lattices with confidence scores calculated as in [28]. Words in the reference that occur in the lattices are given a high score to force the selection of that word. A sim-

ilar approach was used by Van Dalen et al. [7] for hypotheses from crowd sourced data. Manohar et al. [16] propose a different approach that combines four transcripts into a confusion network for training. In Chen et al. [14] the authors align the transcription to a sausage lattice version of the hypotheses (*consensus network*), and select words at each arc depending on the posterior probabilites of the words and the match with the aligned word. Venkataraman et al. [13] proposed to align the data with a robust alignment procedure, using transducers that allow for words to be skipped, and certain insertions, for which transition probabilities are estimated empirically on held-out data. The resulting best path is used as training data. A related approach was taken by Nicolao et al. [8], in which they propose to use a transducer to model variations in children's speech. In Olcoz et al. [12] the authors propose to correct for word-boundary errors and insertions in a lightly supervised alignment. They compare alignments and their confidences from different acoustic models, and include a model specifically trained to detect insertions.

Most of the aforementioned techniques use a biased LM. This is typically a large background LM which is interpolated (using a high weight) towards an LM estimated on the domain of the transcriptions (e.g. [2]), or just the transcriptions themselves (e.g. [4]).

3. Semi-supervised LF-MMI

The LF-MMI objective was introduced by Povey et al. [18] as a method to train neural networks discriminatively with the MMI criterion, without requiring a first-pass of cross-entropy training. It was later extended to the semi-supervised training scenario [19], and to test-time adaptation [20]. We note briefly that the idea of semi-supervised LF-MMI is to generate supervision lattices by decoding unsupervised data using a seed model. Semi-supervised training, and LF-MMI, are well known, and we refer to the above papers for more details.

4. Lattice combination

We present a new method to combine lattices with inaccurate transcripts. The lattices will typically be the output of a decode with a biased LM, and the transcriptions are in this paper subtitles from broadcast data. As in Long et al. [4], we make the assumption that if a word in the transcription is present in the lattice, then the word is likely correct; but also that a transcript word not occurring in the lattices is wrong. Seen from the view of the transcriptions, then, we would like to insert, and substitute, with hypotheses from the lattice when this is the case. Viewed from the lattice, we want to collapse the lattice onto words in the transcription when possible. This provides a narrow lattice where we believe it should be confident, and wide otherwise. Consequently, if the transcriptions are not at all present in the lattice, then the lattice is kept in its entirety. An example output of the algorithm is shown in Figure 1.

To perform the combination, we require a linear transducer, R, of a transcription, and a hypothesis lattice H, for a particular utterance (we assume utterance segmentation a priori), projected on words. All weights are scaled to zero. We create an edit transducer, E, that allows for insertions ($\epsilon : w$), deletions ($w : \epsilon$), and substitutions ($w_i : w_j$), and compose them in the following order:

$$T = R \circ E \circ H. \tag{1}$$

As our goal is to maximise the number of correct words in the lattice, not to minimise the number of edits, we do not use the standard costs for E. Instead, we set every edit cost to 0, apart from matches ($w_k : w_k$) which we set to -1. The path with the most correct words will then have the most negative total cost. We retain only the paths with that minimum cost, or some multiple of it, by pruning the transducer with a threshold t times the shortest path cost:

$$t \otimes [\oplus_\pi w(\pi)], \tag{2}$$

where the sum is over all paths π in $R \circ E \circ H$. We set $t = \bar{0}$ in our experiments.

To obtain the final combined lattice, we first project the pruned transducer on the output. This retains any substitutions made by the hypothesis lattice, and deletes words in R that were not present in H. Finally we remove epsilons, determinise, and minimise. The complete set of operations are:

$$T = \min(\det(\text{rmeps}(\text{proj}(\text{prune}(R \circ E \circ H))))). \tag{3}$$

To use the combined transducer T as supervision, we add back costs by composing it with G, and use the resulting grammar to compile new training graphs with which we align the data to add acoustic costs.

The edit transducer E is not particularly efficient in that it needs to store $(|V| + 1)^2 - 1$ transitions for a vocabulary V, and the resulting search space is quadratic in the length of the input. This could be mitigated with factored transducers, three-way compositions or using a rho-matcher. We note, however, that the computation time spent during lattice combination is negligible compared to the decode pass required to create H.

Shown in Table 1 are the expected WERs over the lattices with respect to the true transcriptions in the test set. The table indicates a considerable improvement with the combined lattices, and demonstrates the inaccuracy of the transcriptions.

Table 1: *Expected WERs over lattices with aligned and segmented references on the test set of Scottish Parliament data.*

Supervision	$\mathbb{E}$[WER]
Transcriptions (R)	55.3
Decode-biased (H)	35.5
Combined-biased (T)	26.4

5. Experimental setup

The baseline model in this paper is trained on news data from the Multi Genre Broadcast (MGB) corpus [1]. The news data is filtered using MER 40% with respect to the lightly supervised decode that is included in the MGB challenge. The resulting data consists of 179 hours across 545 shows, and approximately 15,600 (unlinked) speakers.

We use 5 hours of data from the Scottish Parliament[1] as adaptation data. This consists of 1300 utterances across 374 speakers. On average the utterances contain 30 words each. The test set contains 6.8 hours of audio across 40 speakers. The accompanying subtitles are inaccurate, as demonstrated in Table 1. Empirically we observed large amounts of paraphrasing in the data.

[1] `https://www.youtube.com/user/ScottishParl`

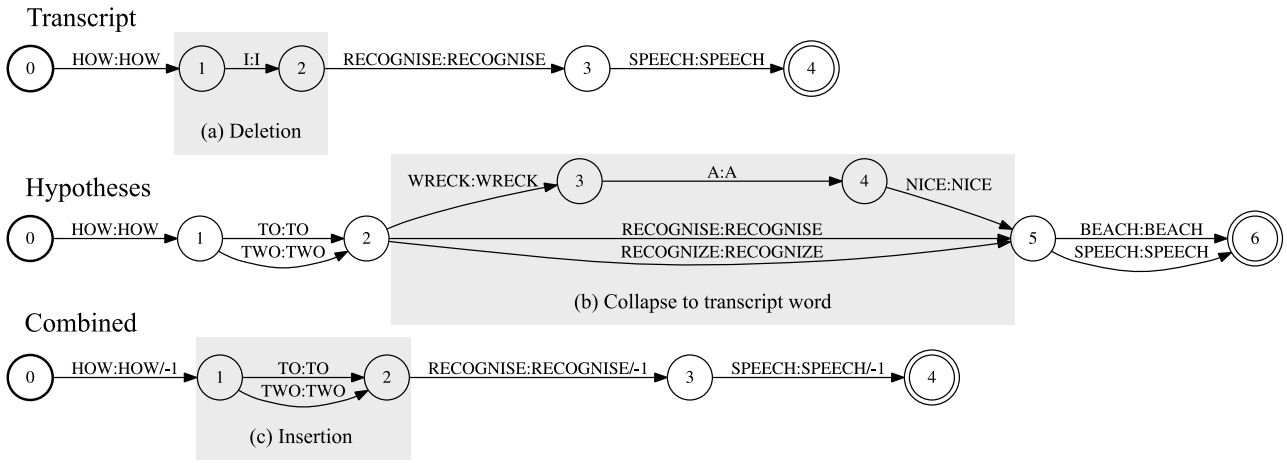

Figure 1: *An example of the result of the proposed lattice combination algorithm. (a) A transcript word is deleted where it is not present in the lattice; (b) the lattice is collapsed where there is a match; (c) insertions missing in the transcriptions are kept or inserted.*

5.1. Model

The baseline model is trained using Kaldi [29], and is based upon the 7p recipe for Switchboard. This is a factored time-delay neural network (TDNN-F) model [30] with 12 layers, each with 1280 units (apart from the penultimate layer), and bottleneck dimensions of 256. Interleaving the layers are ReLU activations, batchnorm and dropout layers. We train on alignments obtained from a standard HMM-GMM system that matches the parameters set out in the MGB challenge [1]. The model is trained with speed-perturbed [31] MGB news data for 8 epochs. The background trigram LM is trained on 640 million words of BBC subtitle text, and is restricted to the top 150,000 unigrams. We estimate biased LMs on the adaptation data in the same way, interpolating with a weight of 0.7. All models are evaluated with the background LM. During semi-supervised training, we train for 3 epochs with an initial learning rate of 5×10^{-5}. The lattice combination is implemented using Kaldi [29] and OpenFST [32].

6. Experiments

Baseline results adapting to the raw transcriptions or in a semi-supervised manner are shown in Section 6.1. Results with the lattice combination are presented in Section 6.2, and with biased LMs in Section 6.3. We experiment with an alternative method to control for deletions in the supervision in Section 6.4.

6.1. Baseline model

The baseline model trained on BBC news data achieves 30.0% WER on the Scottish Parliament test data, as shown in Table 2. Adapting using the unfiltered transcriptions as supervision increases the error rate to 33.2%. This is expected given the high error rate of the transcriptions with the true reference (Table 1). The large proportion of deletion errors suggests that the transcriptions have failed to account for words present in the audio.

In contrast, adapting in a semi-supervised manner, having generated supervision with a decode of the data, improves results. Primarily the number of deletions have dropped, which indicates that the semi-supervised supervision has filled in deletions that were absent in the transcriptions. We also note that training on the best path is worse than using the entire lattice, which is consistent with the literature [19, 20].

Table 2: *Baseline results on Scottish parliament data.*

Method	WER %	Sub	Del	Ins
Baseline	30.0	15.8	11.3	3.0
Transcriptions	33.2	11.0	20.5	1.7
Semisup	28.6	11.7	14.7	2.1
Semisup-BP	28.8	12.9	13.8	2.1

6.2. Lattice combination

Table 3 shows the results of the lattice combination method compared with purely semi-supervised approaches. The combined approach reduces WERs up to 17.5% relative to `Semisup`, with a substantial drop in deletion and some substitution errors. As discussed above, the key difference between the combined supervision and the standard approach is that the decoded lattices typically contain multiple confusable hypotheses where they match the transcriptions, while the combined lattices will have a lattice depth of 1 in these cases. This is reflected in the average lattice depth of the supervision lattices: 21.2 for the combined lattices compared to 78.1 in the original hypotheses. Additionally, the gap between best path and lattice supervision for the combined approach is larger, suggesting that it is benefiting from uncertainty encoded by the wide lattices where it was joined with the original decode.

Table 3: *Combination and semi-supervised results.*

Method	WER %	Sub	Del	Ins
Baseline	30.0	15.8	11.3	3.0
Combined	23.6	10.7	10.7	2.3
Combined-BP	25.2	12.2	10.2	2.7
Semisup	28.6	11.7	14.7	2.1
Semisup-BP	28.8	12.9	13.8	2.1

6.3. Biased language model

The results in Table 4 demonstrate the benefit of including a biased LM when generating lattice supervision. For both methods it reduces the WER by up to 10% relative, compared to Table 3, benefiting both lattice and best path supervision. The standard semi-supervised method seems to benefit more from biasing the LM than the combined method. In future work we would like to investigate whether this effect diminishes with a large in-domain (Scottish Parliament) LM, for which biasing will be less impactful.

Table 4: *Results with an LM biased to the adaptation data.*

Method	WER %	Sub	Del	Ins
Baseline	30.0	15.8	11.3	3.0
Combined-biased	23.3	10.8	10.2	2.4
Combined-biased-BP	25.0	12.0	10.2	2.8
Semisup-biased	26.8	11.7	13.1	2.0
Semisup-biased-BP	26.6	12.0	12.5	2.1

6.4. Controlling for deletions when generating supervision

The results shown thus far indicate that the seed model is inclined to delete, which has affected the generation of supervision lattices. Deletion errors account for more than half of the errors when using `Semisup` supervision. In contrast, the `Combined` method appears to control for deletions. We considered whether there is another option to compensate for a tendency to delete. We propose to achieve this by penalising deletions (or rewarding insertions) in the HCLG[2] decoding graph when generating the lattices. This is implemented by subtracting a constant from every word output label in the graph. We note that a standard deletion penalty is no longer current practice, as a correctly tuned system does not require it. Indeed, we did not find a penalty on the final decode to be helpful. What we are proposing is instead to include a penalty for a specific type of training, and crucially only during the generation of supervision. We decode the final model in the standard fashion.

The effect of including a deletion penalty is shown in Figure 2. All models benefit strongly from the penalty. The detailed results with the optimal penalty for our experiment are shown in Table 5. By including this penalty for supervision generation, the WERs after adaptation drop, along with the number of deletion errors, by up to 13% relative. The combined method now yields 22.8% WER, a small improvement upon the previous result (Table 4), and the best result we obtained in this paper. It still improves upon standard semi-supervised training, but the difference is now less. However, tuning the deletion penalty hyperparameter is time-consuming, as it requires entire passes of decoding to generate supervision. The non-penalised combined result is close to the best overall result.

With a deletion penalty of 3 the lattices grow very large, increasing disk usage and the time to generate training examples, by several orders of magnitude. This is reflected in the average lattice depths, which for the semi-supervised lattices is now 418.0, compared to 10.7 for the combined lattices. Note that the lattice depth for the combined lattices has actually reduced, since the transcriptions are likely to match to more words, as more words are present in the hypothesis lattices.

[2]HCLG denotes the composition of the following WFSTs: acoustic HMM (H), context (C), lexicon (L), and language model (G).

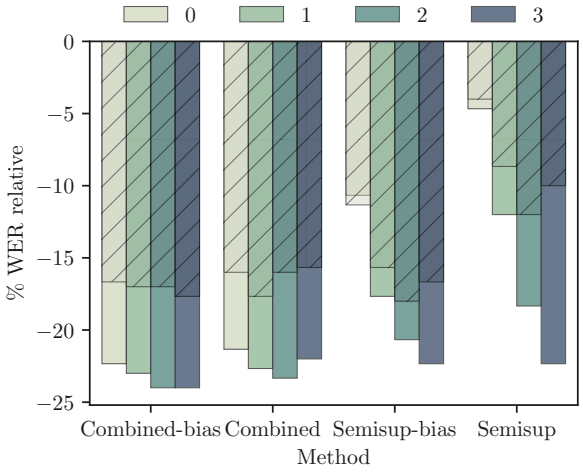

Figure 2: *Results with respect to the baseline (30%) for increasing deletion penalty / insertion reward when generating supervision lattices. The overlaying, hashed, bars represent the performance when using the corresponding best path supervision.*

Table 5: *Results using a deletion penalty of 3 when generating supervision lattices. Transcriptions are not affected.*

Method	WER %	Sub	Del	Ins
Baseline	30.0	15.8	11.3	3.0
Transcriptions	33.2	11.0	20.5	1.7
Combined-biased	22.8	10.9	9.0	2.9
Combined-biased-BP	24.7	11.9	9.9	2.8
Combined	23.4	12.1	7.8	3.5
Combined-BP	25.3	12.9	9.0	3.4
Semisup-biased	23.3	11.2	9.4	2.7
Semisup-biased-BP	25.0	12.8	8.4	3.9
Semisup	25.8	12.7	9.6	3.5
Semisup-BP	27.0	13.8	9.1	4.1

7. Conclusions and future work

We proposed a method for lightly supervised training, to combine inaccurate transcriptions with the decoded hypothesis lattices of a seed model. This produced an improvement upon a purely semi-supervised approach by up to 17.5% relative. Biasing the background language model to the data was found to substantially improve both the semi-supervised training, and the lattice combination technique. We finally suggested a deletion penalty used during the generation of the hypothesis lattices, which was effective when using a seed model that was prone to delete. The combined use of the above ideas produced a WER reduction of 20% with respect to the semi-supervised result.

In future work we would like to investigate improvements to the combination algorithm, how the use of stronger in- and out-of-domain language models affect the results, the effect of the pruning threshold t, and the above on larger amounts of data.

8. Acknowledgements

This work was partially supported by a PhD studentship funded by Bloomberg, and by the EU H2020 projects SUMMA (grant agreement 688139) and ELG (grant agreement 825627).

9. References

[1] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, and P. C. Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 687–693.

[2] J. Driesen and S. Renals, "Lightly supervised automatic subtitling of weather forecasts," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 452–457.

[3] D. Graff, "An overview of broadcast news corpora," *Speech Communication*, vol. 37, no. 1-2, pp. 15–26, 2002.

[4] Y. Long, M. J. Gales, P. K. Lanchantin, X. Liu, M. S. Seigel, and P. C. Woodland, "Improving lightly supervised training for broadcast transcription," in *Interspeech*, 2013, pp. 2187–2191.

[5] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2013, pp. 368–373.

[6] L. Mathias, G. Yegnanarayanan, and J. Fritsch, "Discriminative training of acoustic models applied to domains with unreliable transcripts," in *2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 2005, pp. I–109.

[7] R. C. Van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4709–4713.

[8] M. Nicolao, M. Sanders, and T. Hain, "Improved acoustic modelling for automatic literacy assessment of children," in *Proceedings of Interspeech 2018*. ISCA, 2018, pp. 1666–1670.

[9] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

[10] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–737.

[11] S. Li, Y. Akita, and T. Kawahara, "Discriminative data selection for lightly supervised training of acoustic model using closed caption texts," in *Interspeech*, 2015, pp. 3526–3530.

[12] J. Olcoz, O. Saz, and T. Hain, "Error correction in lightly supervised alignment of broadcast subtitles," in *Interspeech*, 2016, pp. 2110–2114.

[13] A. Venkataraman, A. Stolcke, W. Wang, D. Vergyri, J. Zheng, and V. R. R. Gadde, "An efficient repair procedure for quick transcriptions," in *Interspeech*, 2004, pp. 1961–1964.

[14] L. Chen, L. Lamel, and J.-L. Gauvain, "Lightly supervised acoustic model training using consensus networks," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–189.

[15] O. Saz, S. Deena, M. Doulaty, M. Hasan, B. Khaliq, R. Milner, R. W. Ng, J. Olcoz, and T. Hain, "Lightly supervised alignment of subtitles on multi-genre broadcasts," *Multimedia Tools and Applications*, vol. 77, no. 23, pp. 30 533–30 550, 2018.

[16] V. Manohar, D. Povey, and S. Khudanpur, "JHU Kaldi system for Arabic MGB-3 ASR challenge using diarization, audio-transcript alignment and transfer learning," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 346–352.

[17] A. Stan, P. Bell, and S. King, "A grapheme-based method for automatic alignment of speech and text data," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 286–290.

[18] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016, pp. 2751–2755.

[19] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4844–4848.

[20] O. Klejch, J. Fainberg, P. Bell, and S. Renals, "Lattice-based unsupervised test-time adaptation of neural network acoustic models," *arXiv preprint arXiv:1906.11521*, 2019.

[21] K. Yu, M. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[22] P. Lanchantin, M. J. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P. C. Woodland, and C. Zhang, "Selection of multi-genre broadcast data for the training of automatic speech recognition systems." in *Interspeech*, 2016, pp. 3057–3061.

[23] N. Braunschweiler, M. J. F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Interspeech*, 2010, pp. 2222–2225.

[24] L. Nguyen and B. Xiang, "Light supervision in acoustic model training," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–185.

[25] S. Li, Y. Akita, and T. Kawahara, "Automatic lecture transcription based on discriminative data selection for lightly supervised acoustic model training," *IEICE TRANSACTIONS on Information and Systems*, vol. 98, no. 8, pp. 1545–1552, 2015.

[26] P. Bell and S. Renals, "A system for automatic alignment of broadcast media captions using weighted finite-state transducers," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 675–680.

[27] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (ASRU)*. IEEE, 1997, pp. 347–354.

[28] M. S. Seigel and P. C. Woodland, "Combining information sources for confidence estimation with CRF models," in *Interspeech*, 2011, pp. 905–908.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[30] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," *Interspeech*, pp. 3743–3747, 2018.

[31] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015, pp. 3586–3589.

[32] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *International Conference on Implementation and Application of Automata*. Springer, 2007, pp. 11–23.